ADVANCES IN APPLIED MATHEMATICS 9, 7–21 (1988)

Pattern Analysis of the Genetic Code

MARCELA D. PERLWITZ*

Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113

CHRISTIAN BURKS

Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, University of California, Los Alamos, New Mexico 87545

AND

MICHAEL S. WATERMAN*

Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113

The genetic code is examined in a new and systematic fashion: we consider the code as mapping of one finite set (the 64 codons) to another (the 20 amino acids). Given a class of mappings simpler than the actual code, we ask which mappings best approximate it. This leads to an analysis of the effects of ambiguities (codon degeneracy) in one or two positions. With the 0–1 metric (counting the amino acids as equal or not equal), the codon third base degeneracy is apparent, but the first and second positions are indistinguishable; with the integrated amino acid "distance" metric compiled by Sneath (*J. Theoret. Biol.* **12** (1966), 157–195), the analysis ranks the information content of the three codon positions as follows: 2nd > 1st > 3rd. We discuss possible further applications of this approach to patterns in the genetic code and other codes.

1. INTRODUCTION

There has been interest in characterizing patterns in the genetic code since it was first elucidated: the grouping of codons that code for the same amino acid is an obvious example [3]; a second example is the grouping of codons that code for similar (but not identical) amino acids. Here, similarity is usually defined in terms of one or more chemical characteristics of the amino acids (e.g., hydrophobicity). Several methods, some more formal

*Supported by the System Development Foundation.

0196-8858/88 $7.50

than others, for analyzing these patterns have appeared in the literature. For representative examples, see Alff-Steinberger [1], Crick [4], Goldberg and Wittes [6], Jungck [9], McKay [12], Pelc and Welton [13], Swanson [16], Volkenstein [17], and Woese [18]; several of these papers are reprinted in Jungck [10]. From early on (see [4, 7]) up to the present (see the collection of papers in [10] and the recent review by Jukes [8]), proposed patterns in the genetic code have been a cornerstone for many of the arguments concerning evolution of the code. This work has taken on a new dimension with the recent discovery that mitochondria employ genetic codes that differ from the "universal" code (see the review by Cedergren [20]).

In this paper we examine the genetic code in a new and systematic way. The genetic code is formally viewed as mapping of the 64 codons (triplets of RNA bases) to the 20 amino acids and the termination operator (Table I) illustrates the "universal" genetic code). There are $21^{64} \approx 4 \times 10^{84}$ possible mappings. Though it is impossible to examine all of these maps, we do select a subset relevant to current questions concerning the genetic code and analyze them in detail.

In particular, our view of the genetic code as a mapping of one finite set to another enables us to ask which mappings, selected from a set of mappings that are simpler than the actual genetic code, best approximate the genetic code. Development of these classes of mappings is motivated by the fact that codons are base triplets, and it formalizes the observation that certain bases in various codon positions are equivalent (or almost equivalent) for the amino acid specified. An unbiased assessment of wobble [3] results from this analysis.

Our first analysis, counting amino acids as equal or not equal, shows the 3rd base degeneracy but does not distinguish between the first two bases. Biological "folklore" ranks codon position importance as follows: 2nd > 1st > 3rd. Our analysis using an amino acid distance metric gives this same ranking. This is the first analytical derivation of the ranking. Beyond this application to the genetic code itself, the analysis provides a generalized approach for finding patterns in mappings of finite sets.

2. Analysis of the Genetic Code

2.1. *Patterns in the Genetic Code*

Viewed abstractly, the genetic code is a language in which 64 possible combinations of the four bases—uracil (U), cytosine (C), adenine (A), and guanine (G)—taken three at a time specify either a single amino acid or peptide chain termination. With 64 possible "words" and 21 possible "meanings," there is clearly the potential for different codons coding for

TABLE I
Genetic Code, Consisting of 64 Triplets and Their Corresponding Amino Acid, Is Shown in Its Most Common Representation

2nd	U	C	A	G	3rd
1st	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	TC	TC	A
	Leu	Ser	TC	Try	G
	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note. The three codons marked TC are termination signals of the polypeptide chain. The codes are

Ala = alanine	Leu = leucine
Arg = arginine	Lys = lysine
Asp = aspartic acid	Met = methionine
Asn = asparginine	Phe = phenylalanine
Cys = cysteine	Pro = proline
Glu = glutamic acid	Ser = serine
Gln = glutamine	Thr = threonine
Gly = glycine	Try = tryptophan
His = histidine	Tyr = tyrosine
Ile = isoleucine	Val = valine

identical amino acids. This is in fact the case: many pairs of codons that differ only in the third position base code for the same amino acid. On the other hand, a pair of codons differing only in the first or second position most often code for different amino acids.

Crick [3] proposed a unified model, called the "wobble hypothesis," for all codon-amino acid ambiguities, based on ambivalent interactions between codons and anticodons (the RNA unit that mediates the transfer of

information from the codon to the ribosome that polymerizes the amino acids). In this model, codon–anticodon interactions for the first and second positions are based on the normal Watson–Crick base pairing. However, codon–anticodon interactions for the third codon position, although involving base-pairing, are such that the anticodon base can alternatively pair with more than one base. For example, *G* can pair with *U* as well as *C*. By invoking a class of these atypical base pairs, the model accounts for the observed degeneracies in the genetic code. The model is supported by the observed specificities of trinucleotides binding to tRNA's (Soll *et al.* [15]).

While the wobble hypothesis is a well-accepted explanation of the third-position degeneracy, other aspects of the patterns apparent in the genetic code have not been as amenable to analysis. There have been many proposed groupings of codons based on various chemical characteristics of the amino acids they code for; for instance, a correlation between amino acid chemical structure (hydrophobicity) and the base in the second position of the codon has been proposed by several groups [18, 13, 17].

Alff-Steinberger [1] analyzed the code from the point of view of determining whether or not it was an error minimizing code. He compared the error transmitting property of the universal code with that of computer-generated random codes, considering amino acids properties such as molecular weight, polar requirement [19], and number of dissociating groups. He concluded that for the universal genetic code, single-base substitution in the first position of the codon tends to result in the substitution of an amino acid more similar to the original amino acid than would be expected from a random code; the second position plays the largest role in determining the properties of the amino acid.

Alff-Steinberger's work is the closest in the literature to our own. He compares the effects of single base substitutions in the first or second codon position on the resulting amino acid. These computations are performed for the "universal" code and a number of random codes. Our own work quite simply looks for position-dependent patterns in the "universal" code. This basic analysis has not been previously performed, perhaps due to difficulties in definition and computation. See also Ehrenfeucht *et al.* [5] for some related mathematical work on pattern recognition for functions of several variables. Other more sophisticated analyses have, interestingly enough, been performed. Some of these rest on coding theory [16] and on group theory [2].

2.2. *The Genetic Code as a Mapping*

The biological context of protein synthesis suggests consideration of the genetic code as a mapping which relates two specific sets, the set of trinucleotides (codons) and the set of amino acids (and the termination

signal). This formalism should provide a useful basis for answering questions about correlations between patterns in the code and the structural/functional properties of both the cellular machinery for synthesizing proteins and the proteins themselves. We seek to define sets of mappings that account for patterns in the genetic code. Mappings can be defined that allow evaluation of ambiguities in one or more positions. Here we analyze the effects of ambiguities in one or two positions; the results are compared with known effects of single base substitution.

2.3. *General Description and Notation*

Let $\mathbf{N} = \{A, C, G, U\}$ be the set of nucleic acids, $\mathbf{C} = \{(x_1x_2x_3): x_1, x_2, x_3 \in \mathbf{N}\}$ be the set of 64 codons and **A** be the set of the 20 amino acids and the termination codon TC.

Let the mapping g: $\mathbf{C} \to \mathbf{A}$ be such that it assigns to each triplet in **C** an amino acid in **A** exactly as shown in Table I. That is, g is the usual genetic code. For example, $g(\text{ACG}) = \text{Thr}$ while $g(\text{UGG}) = \text{Try}$. g is simply a function with domain **C** and range **B**.

As can be seen in Table I, any two triplets having the first two bases in common and whose third base is either U or C, code for the same amino acid. This leads to consideration of a collection of 16 disjoint subsets of **C**:

$$X_1 = \{\text{UUU}, \text{UUC}\}, \qquad X_2 = \{\text{UCU}, \text{UCC}\},$$
$$X_3 = \{\text{UAU}, \text{UAC}\} \cdots X_{16} = \{\text{GGU}, \text{GGC}\}.$$

To describe this behavior of the genetic code in mathematical terms, we notice that for all i, $g(x) = g(y)$ for $x, y \in X_i$, $i = 1, \ldots, 16$.

Our interest is in finding this and similar patterns in a systematic and quantitative manner. Not all of these patterns hold universally so we introduce a measure which expresses the agreement between a pattern and the genetic code. First, however, we give a general definition of a pattern, which includes the example just discussed.

DEFINITION 1. Let $\{X_1, X_2, \ldots, X_n\}$ be a collection of disjoint subsets of **C**. For this collection, we define a pattern to be a mapping f satisfying

$$f(x) = f(y) \qquad \text{if } x, y \in X_i$$
$$f(x) = g(x) \qquad \text{if } x \in \bigcup_i X_i.$$

Notice that in the above example

$$\bigcup_{i=1}^{16} X_i = \mathbf{N} \times \mathbf{N} \times \{\mathbf{U}, C\}.$$

For example, X_3 collects the two triplets UAU and UAC that have the first position U, the second position A, and a U or C in the third. A short way to denote this collection of sets is $\{UC\}_3$ indicating that the two first bases of the triplets are any fixed elements of the set **N** but the third base is either U or C. The associated pattern will be $f_{\{U,C\}_3}$, for convenience.

In another example, let $X_i = \{(x_1x_2x_3)\colon x_1 \in \{A,C\} \text{ or } x_3 \in \{U,G\}\}$ for each $x_2 \in \mathbf{N}$. Then $\bigcup_i X_i = \{A,C\} \times \mathbf{N} \times \mathbf{N} \cup \mathbf{N} \times \mathbf{N} \times \{U,G\}$ and the collection will be denoted $\{A,C\}_1 \cup \{U,G\}_3$.

2.4. *Pattern Comparison*

Next, we define the notion of distance between patterns. This is important because we want to find patterns that are "close" to the genetic code.

DEFINITION 2. Let d be a metric on the set **A**. A distance **d** between two patterns f and h will be given by

$$\mathbf{d}(f,h) = \sum_{x \in \mathbf{C}} d(f(x), h(x)). \tag{1}$$

Remark. We remark that **d** is a metric on the set of patterns. For example, take

$$\mathbf{d} = \sum_{x \in \mathbf{C}} \left(1 - \delta_{h,f}(x)\right),$$

where

$$\delta_{h,f}(x) = \begin{cases} 1 & \text{if } f(x) = h(x) \\ 0 & \text{otherwise.} \end{cases}$$

This distance is referred to as "0–1" distance.

Obviously, the smaller the value of **d**, the greater the agreement between the patterns f and h being considered. Our interest is in patterns f with $\mathbf{d}(f, g)$ small. The criteria will be based on the amino acids assigned to the triplets in each particular X_i under the pattern f. Before stating the criteria, we give the following definition.

DEFINITION 3. The cardinality of the domain of a pattern f is defined to be the number of X_i's plus the number of elements in the set $\mathbf{C} - \bigcup_i X_i$.

The wobble hypothesis of Crick motivates this definition. We will denote the domain of f by $D(f)$ and the cardinality of the domain by $|D(f)|$.

It is also useful to consider the amino acids that it is possible to express under a pattern f.

DEFINITION 4. The range of a pattern f, $R(f)$, is defined to be the set of amino acids in the image of f.

The cardinality of the range is equal to the number of elements in $R(f)$ and will be denoted by $|R(f)|$.

2.5. *Criteria for Goodness of Fit of a Pattern*

When assigning amino acids to the triplets of a particular X_i we must make choices for the amino acids. This is done in a way that minimizes the value of $\mathbf{d}(f, g)$ and at the same time maximizes the range of the pattern. When both conditions cannot be satisfied simultaneously, then the condition on the distance prevails. Mathematically it means that we find the pattern f satisfying

$$R(f) = \max\left\{ R(h):\ h \in \left\{ q:\ \mathbf{d}(q, g) = \min_{p} \{\mathbf{d}(p, g)\} \right\} \right\}.$$

EXAMPLES. (1) AUU codes for isoleucine; AUG codes for methionine. Let the collection be $\{\mathrm{U}, \mathrm{G}\}_3$; then $f(\mathrm{AUU}) = f(\mathrm{AUG})$. Therefore the amino acid assigned is either isoleucine or methionine. If we take $\mathbf{d}$ as the 0–1 distance, for either of the two choices, the assignment of amino acids contributes one unit to the value of $\mathbf{d}(f, g)$, but the second choice maximizes the range of f.

(2) Take Example 1 above but with $\{\mathrm{U}, \mathrm{G}, \mathrm{A}\}_3$; then $f(\mathrm{AUU}) = f(\mathrm{AUA}) = f(\mathrm{AUG})$. Since AUA codes isoleucine in the "universal" genetic code, in order to minimize the distance between f and g, the choice is to assign isoleucine to each of the triplets.

2.6. *Pattern Enumeration*

Our interest is in those patterns defined by restricting the alphabet in one or two positions of the codon. By this we mean that the triplets in X_i have two bases in common in the first case and one in the second case. Notice that if the triplets of each X_i have two bases in common, we will have 16 such X_i's and, if the triplets in each X_i have one base in common, we will have four such X_i's.

To determine the number of patterns defined by single positions we consider all possible subsets of N with cardinality 2, 3, and 4. Let $\mathbf{M}$ be one such subset and $X_i = \{(x_1x_2x_3):\ x_k \in \mathbf{M}\}$ with $x_i \in \mathbf{N}$ with $i \neq k$.

If $\mathbf{M}$ has two elements, there are $\binom{4}{2} = 6$ possible subsets. These 6 subsets will determine 6 patterns for each position in the codon. Thus, the total number of these patterns is 18. If $\mathbf{M}$ has three elements, there are $\binom{4}{3} = 4$ possible subsets of $\mathbf{N}$ with three elements each, determining 12 patterns.

Finally, if **M** has four elements (i.e., **M** = **N**), we have three patterns. Therefore, the total number of patterns defined by single positions is 33.

To count the number of patterns defined by double positions, consider two subsets **L** and **M** of the set **N** and let $X_i = \{(x_1x_2x_3\}$: $x_j \in \mathbf{L}$ or $x_k \in \mathbf{M}\}$ for each $x_i \in \mathbf{N}$ with distinct i, j, and k.

If both **L** and **M** have one element each, there are $\binom{4}{1}\binom{4}{1} = 16$ possible ways to combine two such subsets and there are three ways to choose two positions in a triplet, determining 48 patterns. If **L** has one element and **M** has two, we have $2\binom{4}{1}\binom{4}{2} = 48$ combinations of **L** and **M** and therefore 144 patterns are determined. If **L** has one element and **M** has three, there are $2\binom{4}{1}\binom{4}{3} = 32$ ways to choose **L** and **M** and 96 patterns result. Following the same reasoning we get 108 patterns when **L** and **M** have two elements each; 144 patterns when **L** has two elements and **M** has three; 48 patterns when **L** and **M** have three elements each. Notice that all patterns with the same combination of positions, where one of the sets has four elements, are identical. Therefore, we have 3 patterns when **L** has one, two, three, or four elements and **M** has four.

Thus, the total number of distinct patterns defined by two positions in the triplets is 591.

2.7. *Examples*

(1) Let us construct one of the patterns defined by single positions and take the criteria for goodness of fit of a pattern determined by the "0–1" distance as in the example following Definition 2.

Let $X_i = \{(x_1x_2x_3)$: $x_3 \in \{A, G\}\}$ for each pair $x_1, x_2 \in \mathbf{N}$. Then the collection is $\{AG\}_3$ according to the notation established above.

(a) $X_1 = \{UUA, UUG\}$, $X_2 = \{UCA, UCG\}$,
$X_3 = \{UAA, UAG\}$, $X_4 = \{UGA, UGG\}$,
$X_5 = \{CUA, CUG\}$, $X_6 = \{CCA, CCG\}$,
$X_7 = \{CAA, CAG\}$, $X_8 = \{CGA, CGG\}$,
$X_9 = \{AUA, AUG\}$, $X_{10} = \{ACA, ACG\}$,
$X_{11} = \{AAA, AAG\}$, $X_{12} = \{AUA, AUG\}$,
$X_{13} = \{GUA, GUG\}$, $X_{14} = \{GCA, GCG\}$,
$X_{15} = \{GAA, GAG\}$, $X_{16} = \{GGA, GGG\}$.

(b) By definition of a pattern, $f(x) = f(y)$ if $x, y \in X_i$ for all i. Table II shows the amino acid assignment that meets the conditions of the criteria.

(c) Since g(UGA) = TC, f(UGA) = Try, and g(AUA) = Ile, f(AUA) = Met, $\mathbf{d}(f, g) = 2$. That is, Tables I and II differ in two positions.

TABLE II
Amino Acid Assignment Corresponding to the Best Pattern $f_{\{AG\}_3}$

2nd	U	C	A	G	3rd
1st	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	TC	Try	A
	Leu	Ser	TC	Try	G
	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
A	Met	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

(d) Recall that the cardinality of the domain of a pattern was defined to be the number of elements in X_i plus the number of elements in the set $\mathbf{C} - \bigcup_i X_i$. According to this definition we have: 16 elements in $\{X_1, X_2, \ldots, X_{16}\}$ and 32 elements in $\bigcup_i X_i$, so that the cardinality of the domain of $f_{\{A,G\}_3}$ is $16 + (64 - 32) = 48$.

(2) Now, let us construct a pattern defined by double positions. Let $X_i = \{(x_1x_2x_3)\colon x_2 \in \{A, G\}$ or $x_3 \in \{U, C\}\}$ for $i = 1, 2, 3, 4$; the collection will be denoted by $\{A, G\}_2 \cup \{U, C\}_3$.

(a) $X_1 = \{UUU, UUC, UCU, UCC, UAU, UAC, UAA, UAG,$
$UGU, UGC, UGA, UGG\}$
$X_2 = \{CUU, CUC, CCU, CCC, CAU, CAC, CAA, CAG,$
$CGU, CGC, CGA, CGG\}$
$X_3 = \{AUU, AUC, ACU, ACC, AAU, AAC, AAA, AAG,$
$AGU, AGC, AGA, AGG\}$
$X_4 = \{GUU, GUC, GCU, GCC, GAU, GAC, GAA, GAG,$
$GGU, GGC, GGA, GGG\}$.

(b) Table III lists the amino acid assignments.

(c) $\mathbf{d}(f, g) = 35$ because, as shown in Table III, there are 35 cases where the amino acid assignment under f differs from the genetic code.

TABLE III
Amino Acid Assignment Corresponding to the Best Pattern $f_{\{A,G\}_2 \cup \{U,C\}_3}$

2nd	U	C	A	G	3rd
1st	TC	TC	TC	TC	U
	TC	TC	TC	TC	C
U	Leu	Ser	TC	TC	A
	Leu	Ser	TC	TC	G
	Arg	Arg	Arg	Arg	U
	Arg	Arg	Arg	Arg	C
C	Leu	Pro	Arg	Arg	A
	Leu	Pro	Arg	Arg	G
	Lys	Lys	Lys	Lys	U
	Lys	Lys	Lys	Lys	C
A	Met	Thr	Lys	Lys	A
	Met	Thr	Lys	Lys	G
	Gly	Gly	Gly	Gly	U
	Gly	Gly	Gly	Gly	C
G	Val	Ala	Gly	Gly	A
	Val	Ala	Gly	Gly	G

(d) Since we have 4 X_i's with 12 elements each, the cardinality of the domain is $4 + (64 - 48) = 20$.

3. Numerical Results

3.1. As stated before, all patterns we constructed here are defined by restricting the alphabet in one or two positions of the codons. Since the distance between two patterns, by Definition 2, depends on the metric we choose for the set of amino acids, we will first consider the "0–1" distance determined by the metric δ of the example following Definition 2 (i.e., $\mathbf{d}(f(x), h(x)) = \Sigma_{x \in C}(1 - \delta_{h,f}(x))$. Second we choose as the metric for the amino acid that given by Sneath [14] and we will refer to the corresponding distance as the "Sneath distance."

3.2. *Patterns Defined by Single Positions*

The 33 patterns constructed here are those that identify two, three or four nucleotides in one position of the codons. For this reason we expect the patterns to reflect the effects already known about single-base substitution.

TABLE IV
Distances $d(f, g)$ between the Universal Genetic Code and the Maps "f" when Using the "0–1" Distance

		$i = 1$		$i = 2$		$i = 3$	
f	$\lvert D(f)\rvert$	$d(f, g)$	$\lvert R(f)\rvert$	$d(f, g)$	$\lvert R(f)\rvert$	$d(f, g)$	$\lvert R(f)\rvert$
$\{U,C\}_i$	48	14	21	16	21	0	21
$\{U,A\}_i$	48	16	21	16	21	7	21
$\{U,G\}_i$	48	16	21	16	21	8	21
$\{C,A\}_i$	48	14	21	16	21	7	21
$\{C,G\}_i$	48	16	21	16	21	8	21
$\{A,G\}_i$	46	16	21	15	21	2	21
$\{U,C,A\}_i$	32	28	17	32	20	7	21
$\{U,C,G\}_i$	32	30	20	32	21	8	19
$\{U,A,G\}_i$	32	32	19	31	20	9	19
$\{C,A,G\}_i$	34	30	19	31	20	9	20
$\{U,C,A,G\}_i$	16	44	14	47	16	15	15

Note. The size of the domains and ranges are also given.

3.2.1 *The "0–1" Distance*

The main features of the numerical results summarized in Table IV are discussed below:

(1) When identifying the pair of pyrimidines (U and C) in the third position, the corresponding pattern $f_{\{U,C\}_3}$ is exactly the genetic code. Therefore we get $\mathbf{d}(f_{\{U,C\}_3}, g) = 0$ and the assignment is unique which means that there is no possible loss of amino acids in the range.

(2) When identifying the pair of purines (A and G) in the third position, the corresponding patterns differ from the genetic code by two assignments. Thus $\mathbf{d}(f_{\{A,G\}_3}, g) = 2$ and the best of these patterns has the same amino acids as the universal genetic code.

The two features above are frequenctly noted by other authors and are cited and discussed by Crick (1968).

(3) The third position of a codon is less relevant in identifying the amino acid. Indeed the $f_{\mathbf{L}_i}$ is always much smaller for position $i = 3$ for every subset $\mathbf{L}$ of $\mathbf{N}$. This is shown by the above facts and the following: the best of the patterns $f_{\{U,C,A,G\}_3}$ differs from the genetic code by 15 assignments. This is the largest distance from the genetic code among all patterns defined by the third position while the smallest distance between patterns defined by the second position and the genetic code is 15 and the smallest

distance between patterns defined by the first position and the genetic code is 14.

To formalize the relative importance of position over all the patterns, we apply the Friedman test. The Friedman test is a nonparametric test which ranks, for each pattern, the position by relative size (see [11, p. 262]). For example, in Table IV for $f_{\{U,C\}_i}$, the distances for the first row are 14, 16, 0 for $i = 1, 2, 3$. We replace this row by the ranks 2, 3, 1. Then each column is totaled with R_i = sum of column i. We have $k = 3$ columns and $n = 11$ rows. In Table IV, $R_1 = 25.5$, $R_2 = 29.5$, and $R_3 = 11$. To perform a multiple comparisons test, compare the $R_i - R_j$ with $\pm z\sigma$, where $\sigma = \sqrt{nk(k+1)/6} \approx 6.6332$ and z is chosen from the normal table. With an overall level of significance equal to 0.06, each of the 6 comparisons has probability 0.01 and we declare $R_i > R_j$ if $R_i - R_j \geq (2.326)(6.6332) = 15.43$. In this way, $R_1 > R_3$ and $R_2 > R_3$ but no decision is made regarding R_1 and R_2.

It is important to notice that in this analysis the termination codons "TC" are included in the range of each mapping or pattern.

3.2.2. *The "Sneath" Distance*

Before we discuss the numerical results which are summarized in Table V we need to make some important remarks:

(1) Since the "Sneath table" of distances between amino acids (Sneath, 1966) does not include the termination codons, we arbitrarily assign 0 to the distance between any amino acid and the termination signals TC.

TABLE V
Distances between the Universal Genetic Code and the Maps "f" when Using the "Sneath" Distance

f	$i = 1$	$i = 2$	$i = 3$
$\{U,C\}_i$	288	324	0
$\{U,A\}_i$	235	332	116
$\{U,G\}_i$	283	408	176
$\{C,A\}_i$	299	430	117
$\{C,G\}_i$	368	369	176
$\{A,G\}_i$	328	374	22
$\{U,C,A\}_i$	487	648	117
$\{U,C,G\}_i$	576	634	176
$\{U,A,G\}_i$	529	662	176
$\{C,A,G\}_i$	553	679	176
$\{U,C,A,G\}_i$	779	965	294

(2) Because of the nature of the Sneath metric, the resulting minimum distance between the genetic code and the corresponding patterns leads, in most cases, to a unique pattern.

Next we discss the numerical results:

(1) When identifying the pair of pyrimidines (U and C) in the third position, the corresponding pattern $f_{\{U,C\}_3}$ is exactly the genetic code, therefore $d(f_{\{U,C\}_3}, g) = 0$.

(2) The third position is the least relevant among the three. Again, the $i = 3$ column is always much smaller than $i = 1$ or $i = 2$ for each pattern. Also notice the fact that the largest distance between a pattern and the genetic code when restricting the alphabet in the third position is $d(f_{\{U,C,A,G\}_3}, g) = 294$, while the smallest distance between the genetic code and patterns defined by restricting the alphabet in the second position is 324 and the smallest distance between patterns defined by the first position and the genetic code is 235.

(3) The second position is the most sensitive to substitutions and this is reflected by the fact that all the distances between the patterns defined by this position and the genetic code are larger than the corresponding ones defined by the first or third position.

The Friedman test is again applied to analyze the relative importance of $i = 1, 2, 3$. $R_1 = 22$, $R_2 = 33$, and $R_3 = 11$. At what α can we declare $R_2 > R_1 > R_3$? Now again $\sigma \approx 6.6332$ so note that $R_2 - R_1 = R_1 - R_3 = 11$ so that $11 \approx 1.66\sigma$. Thus $\alpha/6 \approx 0.05$ and $\alpha = 0.30$ is necessary to declare $R_2 > R_1 > R_3$. The evidence for this ranking is fairly good but not entirely compelling. A more detailed analysis with double positions in Section 3.3 lends more weight to this ranking.

3.3. *Patterns Defined by Double Positions*

The 591 patterns considered here are those that identify one or more nucleotides in each of two positions of the codon, as illustrated by $f_{\{A,G\}_2 \cup \{U,C\}_3}$ in Section 2.6. The extensive data produced for the "0–1" distance and the Sneath distance is available on request but, in the interest of space, not reproduced here. Instead, the predominant features of the data are studied by the Friedman test. The three columns of the tables correspond to position pairs ($i = 1$, $j = 2$); ($i = 1$, $j = 3$); ($i = 2$, $j = 3$). We have three columns with 197 rows, so that $\sigma = \sqrt{(197)(3)(4)/6} \approx 19.84$. Let R_1 correspond to positions 1 and 2; R_2 correspond to positions 1 and 3; R_3 correspond to positions 2 and 3.

For the "0–1" distance, $R_1 = 590.5$; $R_2 = 238$; $R_3 = 353.5$. (Here ties are resolved by splitting ranks.) The smallest difference is $R_3 - R_2 \approx 5.8\sigma$ which is a significant difference with $\alpha = 2 \times 10^{-8}$. The differences are even more highly significant with the Sneath distance, giving a ranking of $R_1 > R_3 > R_2$. Translating into position pairs, we write

$$\{1,2\} > \{2,3\} > \{1,3\}.$$

Thus we conclude that positions $\{1,2\}$ are the most important in determining amino acids. In fact the ranking of position 2 > position 1 > position 3 is suggested by these results.

4. Summary

The genetic code is formally viewed as a mapping of one finite set (the 64 codons) to another (the 20 amino acids and termination operator). There are $21^{64} \approx 4.19 \times 10^{84}$ possible mappings from a set of 64 to a set of 21 objects. By taking into account the biological setting of our problem, we selected a subset of mappings that are simpler than the universal genetic code. Such a selection is motivated by the fact that the codons are base triplets and that certain bases in various codon positions are equivalent (specify the same amino acid) or almost equivalent (specify amino acids with very similar properties).

To formalize the comparison between the genetic code and any of the mappings chosen, a distance between any two mappings is defined. The resulting distance of course will depend on the metric over the set of amino acids that we choose. In this analysis, only two metrics were considered, the 0–1 metric (counting amino acids as equal or unequal), and the Sneath metric (integrated amino acid distance, compiled by Sneath in 1966). With the 0–1 metric the codon third base degeneracy is significant but the effects of the first and second position are indistinguishable. The Sneath metric shows a significant difference in the effects over each of the three positions, taking them as 2nd > 1st > 3rd, agreeing with the biochemical results.

This study can be extended to consider, individually, various amino acid properties such as water structure former, water structure breaker, mobile electrons, heat and age stability. It will be of interest whether or not the patterns obtained for the Sneath metric hold up for these individual properties. Other obvious areas of interest concern the new mitochondrial codes which have small changes from the "universal" code [20]. These changes in the mitochondrial codes do not seem to fit the general patterns we have deduced for the universal code.

REFERENCES

1. C. ALFF-STEINBERGER, The genetic code and error transmission, *Genetics* **64** (1969), 584–591.
2. M. O. BERTMAN AND J. R. JUNGCK, Group graph of the genetic code, *J. Heredity* **70** (1979), 379–384.
3. F. H. C. CRICK, Codon-anticodon pairing with wobble hypothesis, *J. Molecular Biol.* **19** (1966), 548–555.
4. F. H. C. CRICK, The origin of the genetic code, *J. Molecular Biol.* **38** (1968), 367–379.
5. A. EHRENFEUCHT, J. KAHN, R. MADDUX, AND J. MYCIELSKI, On the dependence of functions on their variables, *J. Combin. Theory Ser. A* **43**, No. 1 (1982), 106–108.
6. A. L. GOLDBERG AND R. E. WITTES, Genetic code: Aspects of organization, *Science* **153** (1965), 420–424.
7. T. H. JUKES, Possibilities for the evolution of the genetic code from a preceding form, *Nature* (1973), 22–36.
8. T. H. JUKES, "Evolution of Genes and Proteins" (M. Nei and R. X. Koehn, Eds.), pp. 191–207, Sinauer Assoc., Sunderland, MA, 1983.
9. J. R. JUNGCK, The genetic code as a periodic table, *J. Molecular Evol.* **11** (1978), 211–224.
10. J. R. JUNGCK, "Origin of the Genetic Code" (J. R. Jungck, Ed.), Dowden, Hutchinson & Ross, London, 1983.
11. E. L. LEHMANN, "Nonparametrics: Statistical Methods Based on Ranks," Holden-Day, San Francisco, 1975.
12. A. L. MCKAY, Optimization of the genetic code, *Nature* **216** (1967), 159–160.
13. S. R. PELC AND M. G. E. WELTON, Stereochemical relationship between coding triplets and amino acids, *Nature* **209** (1966), 868–870.
14. P. H. A. SNEATH, Relations between chemical structure and biological activity in peptides, *J. Theoret. Biol.* **12** (1966), 157–195.
15. D. J. SOLL, E. OHTSUKA, R. D. FAULKNER, R. LOHRMANN, H. HAYATSU, AND F. G. KHORNAN, Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique, *J. Molecular Biol.* **19** (1966), 556–573.
16. R. SWANSON, A vector representation for amino acid sequences, *Bull. Math. Biol.* **46**, No. 2 (1984), 187–203.
17. M. V. VOLKENSTEIN, The genetic coding of protein structure, *Biochem. Biophys. Acta* **119** (1966), 421–424.
18. C. R. WOESE, On the evolution of the genetic code, *Proc. Nat. Acad. Sci. U.S.A.* **54** (1965), 1546–1552.
19. C. R. WOESE, D. H. DUGRE, W. C. SAXINGER, AND S. A. DUGRE, The molecular basis for the genetic code, *Proc. Nat. Acad. Sci. U.S.A.*, **55** (1966), 966–974.
20. R. J. CEDERGREN, An evaluation of mitochondrial transfer-RNA gene evolution and its relation to the genetic code, *Canad. J. Biochem.* **60** (1982), 475–479.